DOCUMENT RESUME

ED 460 146 TM 033 620

AUTHOR Onwuegbuzie, Anthony J.
TITLE A New Proposed Binomial Test of Result Direction.
PUB DATE 2001-11-15
NOTE 30p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (30th, Little Rock, AR, November 14-16, 2001).
PUB TYPE Information Analyses (070) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Effect Size; Error of Measurement; *Hypothesis Testing; *Probability; *Statistical Significance
IDENTIFIERS *Binomial Test; Type I Errors

ABSTRACT

D. Robinson and J. Levin (1997) proposed what they called a two-step procedure for analyzing statistical data in which researchers first evaluate the probability of an observed effect statistically (i.e., statistical significance), and, if and only if, it can be concluded that the underlying finding is too improbable to be due to chance, then they assess its magnitude or effect size (practical significance). This technique is extremely useful when conducting null hypothesis significance tests that have sufficient statistical power. However, if statistical power is lacking, then the first step, which serves as the "gatekeeper" for computing effect sizes, may lead to the nonreporting of a nontrival effect. When multiple tests of statistical significance are conducted, adjustments for inflated Type I error rates should be made to ensure that the actual error rate does not exceed its nominal value. Unfortunately, when multiple tests are undertaken with an adjusted alpha, the statistical power of any particular test is lowered. This paper expands on the Robinson and Levin model by proposing a three-step procedure when five or more hypothesis tests are of interest within the same experiment. In this case, the third step involves the use of a binomial test of result direction (including confidence intervals and effect sizes) to determine whether the number of results falling in a certain direction represents chance by assuming that the probability of any particular result direction is 0.5 under the null hypothesis. (Contains 1 figure, 2 tables; and 85 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running head: BINOMIAL TEST OF RESULT DIRECTION

ED 460 146

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

A. Onwuegbuzie

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

A New Proposed Binomial Test of Result Direction

Anthony J. Onwuegbuzie

Howard University

Correspondence should be addressed to Anthony J. Onwuegbuzie, Department of Human Development and Psychoeducational Studies, School of Education, Howard University, 2441 Fourth Street, NW, Washington, DC 20059, or E-Mail: (tonyonwuegbuzie@aol.com).

TM033620

Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR, November 15, 2001.

BEST COPY AVAILABLE

Abstract

Robinson and Levin (1997) proposed what they called a two-step procedure for analyzing statistical data, whereby researchers first objectively evaluate the probability of an observed effect (i.e., statistical significance) and, if and only if it can be concluded that the underlying finding is too improbable to be due to chance, then they assess its magnitude or effect size (i.e., practical significance). This technique is extremely useful when conducting null hypothesis significance tests that have sufficient statistical power. However, if statistical power is lacking, then the first step, which serves as the "gatekeeper" for computing effect sizes, may lead to the non-reporting of a non-trivial effect. When multiple tests of statistical significance are conducted, adjustments for inflated Type I error rates should be made to ensure that the actual error rate does not exceed its nominal value. Unfortunately, when multiple tests are undertaken with an adjusted alpha, the statistical power of any particular test is lowered. Thus, the present paper expands on Robinson and Levin's model by proposing a three-step procedure when five or more hypotheses tests are of interest within the same experiment. Here, the third step involves the use of a binomial test of result direction (including confidence intervals and effect sizes) to determine whether the number of results falling in a certain direction represents chance, by assuming that the probability of any particular result direction is .5, under the null hypothesis.

A New Proposed Binomial Test of Result Direction

Null hypothesis significance testing (NHST) has been utilized for more than 75 years, stemming from the seminal works of Fisher (1925/1941) and Neyman and Pearson (1928). Despite its widespread use during most of the last century through today, its practice has been extremely controversial. Over the years, several camps have emerged regarding attitudes towards NHST. Opinions stemming from these camps appear to lie on a continuum ranging from those who believe that statistical significance testing should be banned completely (e.g., Bakan, 1966; Cahan, 2000; Carver, 1978, 1993; Cohen, 1994, 1997; Guttman, 1985; Loftus, 1996; Meehl, 1967, 1978; Nix & Barnette, 1998; Rozeboom, 1960; Schmidt, 1992; 1996; Schmidt & Hunter, 1997) to those who contend that statistical significance tests are valid and useful (Abelson, 1997a, 1997b; Frick, 1996; Levin, 1993, 1998; McLean & Ernest, 1998; Mulaik, Raju, & Harshman, 1997; Onwuegbuzie & Daniel, in press-a, in press-b).

Although the debates regarding statistical significance testing have been heated and prolonged, going as far back as Boring (1919) and Berkson (1938, 1942), most researchers agree that practical significance should play a role in the interpretation of empirical data. That is, many analysts agree that effect sizes measures should be provided. However, differences lie with respect to the exact role of effect sizes in research reports. Whereas some researchers (e.g., Thompson, 1996) advocate that effect sizes always be reported regardless of whether statistical significance is found, some (e.g., Carver, 1993) contend that effect sizes should replace statistical significance testing completely. Still others (e.g., Fan, 2001) assert that statistical significance testing (i.e., *p*-values) and effect sizes should complement one another. To date, the latter seems to be the most popular.

Of those who believe that *p*-values and effect sizes should be combined in a meaningful manner, the technique advocated by Robinson and Levinson (1997) appears to be the most structured. Specifically, these researchers proposed what they termed a "two-step process" (p. 22) for testing hypotheses. According to this model, a statistical significant observed finding is followed by one or more indices of practical significance; however, no effect sizes are reported in light of a non statistical significant finding. In other words, researchers should determine first whether the observed effect was statistically significant (i.e., a relationship or a difference that exceeds chance) (Step 1), and, if and only if statistical significance is found, then they should report how large or important the observed finding is (Step 2). As such, the statistical significance test in Step 1 serves as a gatekeeper for the reporting of effect sizes (i.e., practical significance) in Step 2.

The rationale for Robinson and Levin's (1997) two-step model stems from their desire to prevent the over-interpretation of large effect sizes "in the absence of formal assessments of their likelihood" (p. 23). Robinson and Levinson noted the following:

> The main point here, however, is that although effect sizes speak loads about the magnitude of a difference or relationship, they are, in and of themselves, silent with respect to the probability that the estimated difference or relationship is due to chance (sampling error). Permitting authors to promote and publish

> seemingly 'interesting' or 'unusual' outcomes when it can be documented that such outcomes are not really that unusual would open the publication floodgates to chance occurrences and other strange phenomena. So, what is our bottom-line conclusion? Researchers cannot live by effect sizes alone! Just as they should not get by exclusively with *p*s, neither should they get by exclusively with Cohen *d*s and Pearson *r*s. And the bottom-line, two-step editorial policy recommendation? First convince us that a finding is *not due to chance*, and only then, assess how *impressive* it is. (p. 25) [italics in original]

A valid criticism of Robinson and Levin's (1997) two-step model was posited by Cahan (2000). Cahan argued that the inferential statistical test and the corresponding effect are technically unrelated procedures. According to Cahan, the two-step approach is valid only if the statistical significance or non-significance of observed effects provides information about the amount of error (e_s) included. Cahan notes that (a) statistically non-significant effects do not necessarily consist entirely or mostly of noise, and can be error-free; and (b) statistically significant effects are not necessarily error-free, nor do they necessarily contain a trivial amount of random error. Thus they are not necessarily "real," and can be due mostly to error. As such, Cahan concludes that the logic of the two-step approach, which assumes a strong negative relationship between the statistical significance of observed effects and the amount of random error contained, is flawed.

Levin and Robinson (2000) countered Cahan's (2000) criticism of their two-step model by arguing that null hypothesis significance testing and effect-size estimation *are* both "conceptually and functionally related" (p. 35). Levin and Robinson posit that if researchers secure consistency between hypothesis-testing and effect-size estimation at the analytical and interpretational stages of quantitative studies, they would be fostering what they call "conclusion coherence" (p. 35). According to Levin and Robinson (1999), many examples exist wherein researchers who analyze data with small sample sizes in the absence of *p*-values over-interpret their effect sizes. Indeed, Levin (1998) stated:

> In its extreme form, [such action] degenerates to strong conclusions about differential treatment efficacy that are based on comparing a single score of one participant in one treatment condition with that of another participant in a different condition. (p. 45)

Thus, the purpose of the present essay is fourfold. First, an argument will be made for using null hypothesis tests of significance as a gatekeeper for determining whether effect-size measures should be reported. In particular, the major flaws of relying only on effect-size estimates to assess observed findings will be identified. Other indices will be advocated, including the use of confidence intervals and internal and external replications. Second, it will be shown how Robinson and Levin's (1997) Two-Step approach is inadequate when multiple tests of statistical significance are conducted. Third, a Three-Step Model will be proposed to replace the two-step procedure in cases where multiple tests of statistical significance are of interest. This three-step method involves a Binomial Test of Result Direction at its third stage. A heuristic example will

be provided to illustrate the utility of the three-step approach in such cases. Finally, further applications of the three-step model will be delineated.

Most of the discussion in the first two sections below is not new to the literature or even to my own writing (e.g., Daniel & Onwuegbuzie, 2000; Onwuegbuzie, 2001a, 2001b, in press-a, in press-b, in press-c; Onwuegbuzie & Daniel, in press-a, in press-b; Weems & Onwuegbuzie, in press). However, the fact that the majority of statistics textbooks, including many of the leading ones, still do not present an adequate discussion of the advantages and disadvantages of null hypothesis significance testing, instead typically taking the approach of training students and beginning researchers to be mechanical users of statistical techniques rather than thoughtful, critical users (Darr, 1987), justifies expanded attention to this topic.

What if Only Effect Sizes Were Allowed to be Used to Interpret Statistical Data?

As noted above, several prominent researchers have called for null hypothesis significance testing [NHST} to be banned completely. Indeed, some of the staunchest critics of NHST contend that this practice has been extremely harmful to science. For example, Meehl (1978, p. 817) stated that NHST "is a terrible mistake, a basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." Similarly, Tryon (1998) complained,

> [T]he fact that statistical experts and investigators publishing in the best journals cannot consistently interpret the results of these analyses is extremely disturbing. Seven-two years of education have resulted in minuscule, if any, progress toward correcting this situation. It is difficult to estimate the handicap that widespread, incorrect, and intractable use of a primary data analytic method has on a scientific discipline, but the deleterious effects are doubtless substantial... (p. 796).

Schmidt and Hunter (1997, p. 37) asserted that "Statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution." Another stern rebuke has been provided by Thompson (1992): "This [statistical significance testing] has created considerable damage as regards the cumulation of knowledge" (p. 436). Also, Rozeboom (1997) contended that

> Null-hypothesis significance testing is surely the most boneheaded procedure ever institutionalized in the rote training of science students...[I]t is a sociology-of-science wonderment that this statistical practice has remained so unresponsive to criticism...(p. 335)

Unfortunately, not only are such statements unnecessarily inflammatory, but they are also flawed. These passionate critics of NHST claim that use of this technique represents a lack of scientific rigor. Thus, it is ironic that many of the claims of the staunchest NHST opponents, such as those cited above, lack scientific evidence. Unless empirical evidence (e.g., effect sizes) can be provided for such causal assertions, they should not be made, or at least should be presented as a possibility, rather than as a fact. Isn't that how we tell our students to conduct research and to develop principled arguments? While I agree that statistical significance testing has been misused and I concur with many of the criticisms of null hypothesis tests, it is an

extreme leap to charge that NHST by itself has stunted "the cumulation of knowledge" (Thompson, 1992, p. 436), is "one of the worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817) , or it "retards the growth of scientific knowledge... [and]... never makes a positive contribution" (Schmidt & Hunter, 1997, p. 37). Furthermore, some of the factual claims made in an attempt to invalidate NHST critics either have been accompanied by superficial claims or represent flawed logic. As noted by Krantz (1999),

> It is one thing to accuse scientists of showing their ignorance of statistical reasoning in the course of their science, but this does not imply that their ultimate conclusions will be incorrect, nor even that their efficiency in reaching correct conclusions will be impaired. A causal attribution of this sort needs to be supported by careful empirical arguments.

Nevertheless, many valid criticisms of NHST have been made. Fan (2001) provides an excellent summary of some of these criticisms:

> Thompson (1993) discussed three relevant criticisms on statistical criticisms for statistical significance testing: (a) overdependency on sample size, (b) some nonsensical comparisons, and (c) some inescapable dilemmas created by statistical significance testing (e.g., testing for assumption vs. testing for research hypothesis). In a similar vein, Kirk (1996) discussed three major criticisms of statistical significance testing: (a) significance testing does not tell researchers what they want to know, but rather creates the illusion of probabilistic proof by contradiction (Falk & Greenbaum, 1995), (b) Statistical significance testing is often a trivial exercise because it simply indicates the power of the design (which primarily depends on the sample size) to reject the false null hypothesis, (c) Significance testing "turns a continuum of uncertainty into a dichotomous reject-do-not-reject decision," and this dichotomous decision process may "lead to the anomalous situation in which two researchers obtain identical treatment effects but draw different conclusions" (Kirk, p. 748) because of the slight differences in their design (e.g., sample sizes).
>
> Because of these and other criticisms, many researchers have called for the reporting of effect sizes either to complement or to replace NHST. In fact, very recently, the latest version of the American Psychological Association (APA), version 5, contained the following statement:
>
> When reporting inferential statistics (e.g., *t* tests, *F* tests, and chi-square), include information about the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme as or more extreme than the one obtained, and the direction of the effect. Be sure to include sufficient descriptive statistics (e.g., per-cell sample size, means, correlations, standard deviations) so that the nature of the effect being reported can be understood by the reader and for future meta-analyses. This information is important, even if no significant effect is being reported. (p. 22)

A few pages later, APA (2001) states

> Neither of the two types of probability value directly reflects the magnitude of an

> effect or the strength of a relationship. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. (p. 25)

On the next page, APA states that

> The general principle to be followed, however, is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. (p. 26)

These statements were the direct result of the American Psychological Association (APA) Board of Scientific Affairs (1999), who convened a committee called the APA Task Force on Statistical Inference (TFSI), recommending in no uncertain terms, that researchers should "always present effect sizes for primary outcomes...[and]...reporting and interpreting effect sizes...is essential to good research" (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599).

Thus, it is clear that over the last few decades statistical significance testing has come under extremely close scrutiny. Indeed, since 1950, the number of articles published in the fields of education, psychology, ecology, and medicine criticizing statistical significance testing has been increasing at an exponential rate (Anderson, Burnham, & Thompson, 1999). Additionally, journals (e.g., *Journal of Experimental Education* and *Research in the Schools*) have devoted a special theme issue to NHST; symposia also have occurred at national annual meetings such as the American Educational Research Association, the American Psychological Association, and the American Psychological Society. Even an edited book, entitled *What if there were no significance tests?* (Harlow, Mulaik, & Steiger, 1997), has been devoted to this topic.

Unfortunately, although effect sizes has logical appeal, it is clear that their use have not been subjected to the same level of scrutiny as has NHST. In particular, although the limitations of NHST have been presented by APA (2001), inasmuch as it states that *p*-values do not directly reflect "the magnitude of an effect or the strength of the relationship" (p. 25), no such presentation of the limitations of effect sizes prevails in this pivotal book. As such, it appears that some researchers, if not many, are not aware that many of the same criticisms launched against NHST also can be aimed at effect sizes. In particular, one of the most repeated criticisms of statistical significance testing probably is its overreliance on sample size (Cohen, 1994; Fan, 2001; Kirk, 1996; Onwuegbuzie & Daniel, in press-a, in press-b; Schmidt & Hunter, 1997; Thompson, 1993). Yet, as noted recently by Fan (2001), "effect size can also be misleading because sample size influences the sampling variability of an effect-size measure" (p. 275). Moreover, using Monte Carlo methods, Fan demonstrated that an observed finding that appears to have practical significance (i.e., large effect size) actually could be the result of sampling error, thereby making any resultant conclusions unreliable and potentially misleading. Based on his findings, Fan recommended that information about both statistical significance and effect sizes be reported for observed findings:

> Statistical significance testing and effect size are two related sides that together make a coin; they complement each other but do not substitute for one another. Good research practice requires that, for making sound quantitative decisions in

educational research, both sides should be considered. (p. 275)

Prior to Fan's (2001) work, Barnette and McLean (1999), again using Monte Carlo procedures, found that standardized effect-size variation was systematic rather than random. According to these authors, the number of groups and sample sizes were highly predictive (i.e., R^2 = .999) of standardized effect sizes. Similarly, Hogarty and Kromrey (2001), once again using Monte Carlo methods, demonstrated that the most frequently used effect-size estimates (e.g., Cohen's *1988 d* and Hedges *and Olkin's g*) are extremely sensitive to departures from normality and homogeneity. Even trimmed effect-size measures (Hedges & Olkin, 1985; Yuen, 1974) exhibit extreme bias when the sample is small, as do several non-parametric effect-size indices, such as Y_1 (Kraemer & Andrews, 1982) and the Common Language (CL) effect-size statistic (McGraw & Wong, 1992). Indeed, bearing in mind that different effect-size measures are suitable for different types of data (Hogarty & Kromrey, 2001), it is disturbing that some researchers do not even indicate the index to which they are referring when reporting effect sizes (Kirk, 1996), nor do they appear to indicate whether the effect-size measure interpreted represents an adjusted or unadjusted index. The lack of information provided here is extremely disturbing because meta-analyses are based on aggregating and comparing effect sizes across studies. How can effect sizes be aggregated if it is not clear whether they are measured on the same metric? Unfortunately, the method of some meta-analysts of omitting unlabeled effect sizes from the aggregate index introduces bias.

It should be no surprise that effect sizes are affected in much the same way as are *p*-values by sample size. Indeed, effect-size statistics represent random variables. Consequently, effect-size measures are affected by sampling variability, as dictated by its underlying sampling distribution. In turn, the amount of sampling variability of an effect-size estimate is influenced by the underlying sample size, in much the same way as *p*-values are affected by the number of cases utilized in the study. That is, when the sample size is small, the discrepancy between the sample effect size and population effect size is larger (i.e., large bias) than when the sample size is large. To use an extreme example, should one be excited about obtaining a large correlation coefficient that was based on three cases? Also, effect sizes are affected by non random sampling. Unfortunately, approximately 95% of studies do not involve random sampling. Thus, measures advanced to compensate for the problems stemming from the role of sample size in NHST (e.g., use of confidence intervals) also should apply to effect sizes.

Additionally, support of the findings of Fan (2001), Barnette and McLean (1999), and Hogarty and Kromrey (2001) stems from the fact that many of the effect-size formulae rely heavily on the assumptions of normality and homogeneity of variance being met. For example, Cohen's (1988, p. 20) *d* effect-size measure is given by the following:

$$d = \frac{M_1 - M_2}{\sigma} \quad (1)$$

where σ is the pooled within-population standard deviation (i.e., $\sigma_1 = \sigma_2 = \sigma$). The numerator component of the expression involves means, which are extremely sensitive to extreme observations, especially when sample sizes are small (Huck, 2000). In the small-sample case, an extreme observation in one of the groups (e.g., experimental group) could seriously distort the mean differences, thereby unduly influencing the effect-size estimate. To the extent that the outlying observations affect the *t*-statistic involved in the *t*-test of independent samples (i.e., statistical significance), it also influences the effect size (i.e., practical significance).

Further, because the denominator in Equation 1 utilizes the pooled standard deviation, any heterogeneity of variance would affect the effect size in a similar way to the *t*-statistic. Moreover, the problems caused by departures from normality and heterogeneity of variance when null hypothesis significance testing is involved are very much an issue for effect-size measures associated with more complex family members of the general linear model. For example, the standard effect-size indices (e.g., ϵ^2 and ω^2) that are utilized often for OVA-type (e.g., ANOVA, MANOVA) analyses assume equal variances, which is not always met (Onwuegbuzie & Daniel, in press-b). However, these weaknesses do not imply that effect sizes should be banned or replaced by some other class of indices--which echoes what some researchers (e.g., Carver, 1993) recommend should be the fate of null hypothesis significance testing. Indeed, in cases where such violations come to the fore, non-parametric effect sizes (e.g., Y_1 and CL) may be more appropriate, in much the same way that non-parametric inferential statistics often are more appropriate when normality and/or heterogeneity of variance appears to underlie the data. Rather, the above limitations pertaining to effect sizes identified above suggest that (a) all assumptions underlying the selected effect-size method should be subjected to the same stringent scrutiny as should statistical significance tests, (b) combining null hypothesis significance testing and effect-size indices, after checking all pertinent assumptions, provides an additional safety net from false or misleading conclusions, compared to using either technique alone, and (c) researchers should pay much more attention to maximizing the quality of their research designs (e.g., select an appropriate sample size) in order to minimize threats to the model assumptions that pertain to both the null hypothesis significance test and the accompanying effect-size measure of interest.

In addition, it does not appear to be obvious to some researchers that effect sizes are a function of the scale of measurement used. Evidence of this is provided by McLean, O'Neal, and Barnette (2000), who, in their award-winning paper, demonstrated that gain effect sizes were different for the raw, scaled scores, and Normal Curve Equivalent (NCE) scores for Grades 4, 6, and 8 on a national norm-referenced test for 749, 574, and 464 schools, respectively, representing 120,149 students. Specifically, as McLean et al. expected, the effect sizes for NCE scores were lower than those for raw

and scaled scores. The researchers appropriately concluded that when effect sizes are computed, researchers should take into account the scale of measurement upon which they are being applied.

As noted above, a way in which NHST is abused occurs when a dichotomous decision (i.e., reject vs. do not reject) is used as the sole determinant of the *significance* (i.e., importance) of an observed finding by comparing the *p*-value to .05. Yet, many researchers who interpret effect sizes appear to use religiously criteria such as those provided by Cohen (1988), who popularized the use of effect-size reporting, even though recommendations vary with respect to how effect sizes should be interpreted (McLean et al., 2000), and despite Cohen's (1988) admonishment that effect-size values are dependent on the specific content and methods that prevail in a given research context. For example, in interpreting effect sizes associated with differences between two groups (i.e., Cohen's *d*), Cohen (1988) recommended cutpoints of .20 for small effects, .50 for medium effects, and .80 for large effects. In stark contrast, McLean (1995) suggested the following criteria: .50 for small effects, between .50 and 1.00 for moderate effects, and above 1.00 for large effects. Regardless of which criteria are used, it is clear that adherence to such cutpoints has the effect of trichotomizing interpretations in much the same way as *p*-values dichotomize decision-making. As noted by Shaver (1993), "There already is a tendency to use criteria, such as J. Cohen's (1988) standards for small, medium, and large effect sizes, as mindlessly as has been the practice with the .05 criterion in statistical significance testing" (p. 311).

A valid criticism of NHST that is supported by data pertains to the low statistical power that prevails in many studies. Indeed, the average power of null hypothesis significance tests typically lies between .40 and .60 in empirical studies (Cohen, 1962, 1965, 1988, 1994, 1997; Schmidt, 1996; Sedlmeier & Gigerenzer, 1989). With an estimated point estimate for power across studies of .50. (Cohen, 1962, 1997), as noted by Schmidt and Hunter (1997, p. 40), "This level of accuracy is so low that it could be achieved just by flipping a (unbiased) coin!" Yet, the fact that power is unacceptably low in most studies suggests that misuse of NHST is to blame, not the logic of NHST. Indeed, it can be argued that low statistical power represents more of a research design issue than it is a statistical issue, since it can be rectified by using a larger sample. Unfortunately, as illustrated above, effect sizes also can fall victim to poor research designs in general and small sample sizes in particular. In fact, an obsession with effect sizes without considering the sample size can have the effect of promoting weak research designs. As such, in making decisions about which articles should be published, journal editors should focus less on *p*–values and effect sizes and more on the quality of the underlying research design.

Reliability is a concept that receives disproportionately scant attention in the interpretation of an observed finding (Onwuegbuzie & Daniel, 2000, 2001, in press-a, in press-b; Onwuegbuzie, Daniel, & Roberts, 2001; Roberts & Onwuegbuzie, 2000; Roberts, Onwuegbuzie, & Eby, 2001). Reliability, or rather *unreliability*, can adversely affect the internal validity of findings (i.e., instrumentation; Campbell, 1957; Campbell & Stanley, 1963; Onwuegbuzie, in press-b) through an inflation of Type I error or a

reduction in statistical power. With respect to the latter, Onwuegbuzie and Daniel (2000) demonstrated that subgroups with scores that generate markedly different reliability estimates can seriously reduce statistical power, even when the full-sample reliability coefficients is adequate. However, low reliability indices do not only adversely affect null hypothesis significance testing, but also negatively impact effect-size measures. After all, poor reliability coefficients stem from scores that do not behave in a consistent manner (Onwuegbuzie & Daniel, 2000, 2001), and it is these scores that are used to calculate both test statistics in null hypothesis significance testing and effect-size measures. Thus, effect-size measures are subject to the same limitations stemming from inadequate reliability as are *p*-values. Indeed, effect sizes always should be interpreted with respect to the observed score reliability, as should be the case for statistical significance testing.

Many critics of NHST have advocated that these be replaced by confidence intervals (e.g., Carver, 1993; Schmidt, 1996). A (two-sided) confidence interval, characterized by lower and upper bounds, identifies a probable range of magnitudes for the effect size (Abelson, 1997b). As such, confidence intervals can be viewed as estimates of *theoretical significance*, as opposed to statistical significance (i.e., *p*-values) and practical significance (effect sizes) (Onwuegbuzie, 2001b). Moreover, a confidence interval includes in it all the information provided by the null hypothesis significance test and more (Cohen, 1994; Serlin, 1993). Thus, it is baffling why some researchers recommend that statistical significance testing be replaced by confidence intervals because null hypothesis significance testing can be conducted within the framework of confidence intervals. Once a (95%) confidence interval is constructed, what is to stop an analyst from examining whether 0, or any other theoretical value, is contained in the interval? As noted by Krantz (1999), this is clearly tantamount to statistical significance testing. Thus, replacing NHST with confidence intervals will not prevent researchers from conducting statistical significance tests. Moreover, critics of NHST should not be recommending that confidence intervals replace statistical significance tests but that they should replace the reporting of *p*-values. In any case, the reporting of confidence intervals as indices of theoretical significance has logical appeal because it has the potential to minimize the likelihood of misleading interpretation of large but spurious effect sizes (Cahan, 2000).

In summary, this section has highlighted 10 limitations associated with the use of effect-size indices. Specifically, use of effect sizes is adversely affected by (a) undue sampling error stemming from small sample sizes and non random samples; (b) departures from normality; (c) heterogeneity of variances; (d) unreliability of scores (i.e., measurement errors); (e) lack of information about the effect size measure used, resulting in problems with aggregating different effect-size indices; (f) lack of consideration of scale of measurement; (g) rigid interpretation of effect-size coefficients; (h) multiple (contradictory) cutpoints for effect-size interpretation; (i) multiple interpretations of effect size indices; and (j) problems associated with point estimate nature. Unfortunately, many researchers who criticize null hypothesis significance testing in general and those who advocate replacing *p*-values with effect size measures

in particular often fail to mention any of these limitations associated with the reporting of effect sizes. Thus, analysts who criticize null hypothesis significance testing without also discussing the limitations of effect sizes are not providing a balanced analysis but are focusing solely on the bad practices linked to the former method. These threats to the validity of effect-size reporting make it clear that effect sizes are not a panacea for empirical research. Moreover, the flaws outlined above provide credence to the assertion that null hypothesis significance tests and effect sizes should be used in combination. A logical way of combining these two procedures is via Robinson and Levin's (1997) Two-Step method for analyzing empirical data. It is to this we now turn.

Robinson and Levin's (1997) Two-Step Method

As noted by Onwuegbuzie (in press-b), the overall goal of empirical research is to collect and to analyze data that help the researcher to address research questions and/or to test hypotheses, leading to meaningful conclusions in which as many rival explanations as possible are eliminated. This is the goal that drives both null hypothesis significance testing and effect size reporting. Unfortunately, as was illustrated in the previous section, use of both null hypothesis significance testing and effect sizes contains inherent flaws. Thus, no one index, by itself, is a hegemony in analyzing and interpreting data. Rather, using both methods in combination helps to rule out more rival explanations than would occur if either method was used alone to interpret observed findings.

Among those who agree that *p*-values and effect sizes complement one another, debate exists as to how these indices should be combined. Moreover, controversy prevails about which index should dominate interpretation. Table 1 illustrates the four possible outcomes and their conclusion validity when combining *p*-values and effect sizes. This table echoes the decision table that demonstrates the relationship between Type I error and Type II error. It can be seen from Table 1 that conclusion validity occurs when (a) both statistical significance (e.g., $p < .05$) and a large effect size prevail; and (b) both statistical non-significance (e.g., $p > .05$) and a small effect size occur. Conversely, conclusion invalidity exists if (a) statistical non-significance (e.g., $p > .05$) is combined a large effect size (i.e., Type A error) and (b) statistical significance (e.g., $p < .05$) is combined a small effect size (i.e., Type B error). Both a Type A error and a Type B error suggest that any declaration that the observed finding is meaningful on the part of the researcher would be misleading and hence result in conclusion invalidity. In this respect, conclusion validity is similar to what Levin and Robinson (2000) term *conclusion coherence*. Levin and Robinson define conclusion coherence as "consistency between the hypothesis-testing and estimation phases of the decision-oriented empirical study" (p. 35).

Insert Table 1 about here

Some researchers (e.g. Thompson, 1996) contend that effect sizes always should be reported, that is, regardless of whether statistical significance is observed. In effect, these individuals believe that effect sizes should take precedence over *p*-values

with respect to interpreting statistical data. However, as can be seen in Table 1, this policy may lead to a Type A or Type B error being committed. Similarly, Type A and Type B errors also prevail when null hypothesis significance testing is used by itself. In fact, conclusion validity is maximized only if Robinson and Levin's (1997, p. 22) "two-step" procedure is utilized.

According to this model, null hypothesis significance testing takes precedence over effect-size calculation inasmuch as it represents the first step in the process of analyzing and interpreting statistical data. More specifically, effect sizes are reported if and only if statistical significance is found. Thus, the first step is for the researcher to determine whether the observed finding is statistically real or it represents chance. If the former, then one or more effect-size measure (i.e., practical significance) is reported (i.e., Step 2), because sampling error can be ruled out as a serious rival explanation. If no evidence is found to suggest that the result exceeded what would be expected by chance, no effect sizes are reported. The Two-Step method is consistent with Barnette and McLean (2000) and McLean et al. (2000), who recommended using a statistical significance test before utilizing an effect-size index in order to prevent the interpretation of random effect sizes.

The utility of the two-step process is that it results in a lower probability of Type A and Type B occurring, relative to other combinations of combining information about *p*-values and effect sizes because it prevents researchers from "interpreting descriptive effect sizes as 'real' even though statistical tests of those effects have indicated either that they should be regarded as 'chance' deviations from a null-hypothesis parameter, given sufficient statistical power,...,or that the study's evidence is inconclusive" (Robinson & Levin, 2000, p. 35) (i.e., Type A error) and from "crowing about 'impressive' effect sizes without their having applied any inferential statistical procedures (hypothesis tests or confidence intervals) at all" (p. 350 (Type B error). As such, the Two-Step method appears to have the greatest potential for obtaining conclusion validity or coherence.

Limitations of the Two-Step Procedure

When multiple tests of statistical significance are conducted, as is often the case, adjustments for inflated Type I error rates should be made (Huck, 2000; Onwuegbuzie & Daniel, in press-a, in press-b). For example, when a researcher conducts a single statistical significance test, the probability of rejecting a true null hypothesis is equal to the critical *p* (α)-value for that test. However, when *k* tests of statistical significance are conducted, assuming independence of each test, with testwise error probabilities of α_1, α_2, . . ., α_k for the *k* tests, the *overall alpha* (or "familywise" alpha) level actually exceeds the value of any of the individual testwise alphas:

$$\text{overall } \alpha = 1 - (1 - \alpha_1)(1 - \alpha_2)\ldots(1 - \alpha_k)$$

In the case when all *k* statistical significance tests employ the same testwise alpha level (α_T), the equation could be simplified:

$$\text{overall } \alpha = 1 - (1 - \alpha_T)^k$$

Using this latter formula, in the case of performing 10 tests of statistical significance at the .05 level of probability, the overall Type I error probability would actually be not 5%,

but 40.1%.

In handling this problem, all the tests of statistical significance can be undertaken either by using the same "adjusted" alpha level via techniques such as the Bonferroni adjustment, or by making some tests more liberal than others in the set of correlations analyzed, via methods such as the Holms procedure (Huck, 2000). Whatever technique is used, it is important to attempt to ensure that the actual Type I actual error does not exceed its nominal value. Alternatively, multivariate statistical procedures can be utilized to minimize the number of statistical significance tests employed, thereby countering the Type I error inflation problem, in addition to representing a more realistic picture of the multivariate reality in which the variables actually occur (Onwuegbuzie & Daniel, in press-b).

Although the analysis of multivariate data often honors, in the optimal sense, the nature of reality that most researchers are interested in studying (Tatsuoka, 1973; Thompson, 1999), there are times when focusing on univariate effects is more appropriate. For example, if a child is being diagnosed for learning difficulties, then isolating her/his individual weakness is extremely appropriate. In such cases, researchers should bypass mutlivariate analyses and proceed directly to conducting multiple univariate analyses (Keselman et al., 1998).

Unfortunately, when multiple tests are conducted and an adjusted alpha value utilized, the statistical power of any particular test is lowered in proportion to the number of tests performed. For example, if a researcher wanted to compare 64 males and 64 females on one variable, with an overall rate of .05, the alpha value is .05; if two gender comparisons were made, using the Bonferroni adjustment, the adjusted alpha is .025 (i.e., .05/2); if five gender comparisons are made, then the adjusted alpha value is .01 (.05/5). Thus, the statistical power for detecting a moderate effect will be lower for five comparisons (.59) than for two comparisons (.71) and one comparison (.80). In any case, the statistical power for making more than one comparison is inadequate. Disturbingly, the typical level of power used to detect moderate relationships in the social and behavioral sciences is .5 (Cohen, 1962, 1997). Indeed, many studies are unable to discover true relationships that prevail in the population.

In situations when statistical power is low because of an interest in conducting multiple univariate null hypothesis significance tests, Robinson and Levin's (1997) Two-Step Procedure is problematic. This is because when power is low, the first step of the two-step procedure will lead to an inflation of Type II error. Unfortunately, Type II error and Type A error (cf. Table 1) are positively related. That is, as Type I error increases (i.e., statistical power decreases), Type A error also increases. Thus, the Two-Step method represents an extremely viable technique as long as the statistical significance tests of interests have sufficient statistical power. When power is lower, as is often the case when multiple univariate null hypothesis significance tests are conducted, it is difficult to assess the extent to which the Two-Step method yields conclusion validity because of the increased likelihood of Type A error. As such, when multiple univariate null hypothesis significance tests are of interest, a Three-Step method of analyzing data is proposed.

Three-Step Method for Testing Multiple Univariate Null Hypotheses

As noted above, when conducting multiple univariate tests of significance, it is common to find situations in which after adjusting for Type I error, none of the observed effects were statistically significant, even though at least one of the findings would have been declared statistically significant if the number of hypotheses of interests had been less. In such cases, using the Two-Step method, the researcher would declare that none of the observed findings were statistically significant and thus no effect sizes would be reported. However, such a conclusion can be extremely misleading when low statistical power is a serious rival explanation. For example, let us suppose that a researcher was interested in examining gender differences in levels of statistics anxiety among graduate students. Let us further suppose that this researcher administered the Statistical Anxiety Rating Scale (STARS; Cruise & Wilkins, 1980), which contains six subscales, to a random sample of 100 students (50 females and 50 males). Also, let us assume that she wanted to compare the male and female sample members with respect to these six dimensions individually. (Let us assume that the scores pertaining to all five subscales yielded adequate reliability coefficients.) Unfortunately, after applying the Bonferroni adjustment (i.e., adjusted $\alpha = .008$), the investigator observed that no gender difference emerged for any of the six subscale scores. Using the Two-Step method, the researcher would conclude that no gender differences existed for any of the dimensions of statistics anxiety and she would not report any effect sizes. However, let us suppose that on further investigation, she noted that, although not representing a statistically significant difference, for all six dimensions, females reported higher scores than did males. Is this noteworthy or just a coincidence? Well, for each of the six tests of hypothesis, the null hypothesis is that no gender difference in scores exists. Even if this was perfectly true in the population, because of sampling error, we would expect a difference between the female and male students, however small. Again, assuming no gender difference in the population, we would expect males to report higher scores 50% of the time (on average) and females to report higher scores the other 50% of the time (on average). That is, the expected probability that females (or females) would obtain higher scores under the null hypothesis is .5. Thus, across the six subscales, we would expect males to exhibit higher scores (at random) for three of the subscales, and females to exhibit higher scores (at random) for the other three subscales. However, the researcher found that females reported higher scores for all six subscales. Interestingly, the Binomial distribution can be used to determine the probability of obtaining six outcomes in the same direction (i.e., female students higher on all six occasions), with a per-experiment probability (i.e., for each subscale) of .5. The binomial probability distribution gives the probability associated with each possible x value. Specifically, given a binomial experiment consisting of n trials, where the probability of success on an individual trial is p and the probability of failure is q (where $q = 1 - p$). Then the probability of exactly n successes in n trials is given by the formula

$$p(x) = \frac{n!}{x!(n - x)!} \cdot p^{x} \cdot (1 - p)^{n - x} \qquad x = 0, 1, \ldots, n$$

where $\frac{n!}{x!(n - x)!}$ = *number of outcomes with x successes.*

that is,

$$P(x) = C(n, x)\, p^{x}\, q^{n-x}$$

This formula could be used to determine the probability of obtaining six successes (females reporting higher scores) out of six trials, by setting *p* and *q* equal to 0.5, and *n* = 6, and then letting *x* = 6 (i.e., the number of successes). This yields a probability of .016. Thus, the probability of obtaining the number of successes as extreme as 6 out of 6 is 0.016. Simply put, the probability of obtaining an outcome as extreme as the one observed (i.e., 100% successes) on either side of the binomial distribution based on six null hypothesis tests of significance is only 0.016. Using a 5% level of significance, the researcher would reject the null hypothesis that female students do not consistently report higher levels of statistics anxiety than do the males. (A 95% confidence interval for the proportion of successes also could have been reported.)

For future reference, the test described above will be termed the *Binomial Test of Result Direction*. Because a statistically significant finding emerged, the researcher would complete the analysis by computing an effect-size measure. An effect-size index that is compatible with the Binomial test is the difference in success proportions, which is given by the proportion of successes (i.e., 6/6) minus the proportion of successes under the null hypothesis (i.e., 0.50). For the present data, this index is 1.00 - 0.50 = 0.50, which is the maximum value for this index. An alternative effect-size index for the Binomial test is the proportion-observed-above chance index, which is given by the difference in success proportions divided by the proportion of successes under the null hypothesis (i.e., 0.50), expressed as a percentage. For the current example, the proportion-observed-above chance index is 100(1.00 - 0.50)/0.50 = 100%, which, again is the maximum value for this index. Thus, the researcher could state the following:

> A series of independent *t*-tests, after applying the Bonferroni adjustment, revealed no statistically significant difference between male and female graduate students with respect to scores on all six subscales of the STARS. However, a follow-up Binomial Test of Result Direction revealed that the consistency with which females reported higher anxiety scores was statistically significant. The effect size associated with this trend, as measured by the proportion-observed-above chance index, was maximal.

Alternatively, in reporting the Binomial Test of result Direction, the researcher could have stated that "the trend in anxiety scores were in a statistically real direction."

Figure 1 presents the proposed Three-Step Model for testing multiple univariate null hypotheses. It can be seen from this figure that the first step involves testing for statistical significance each of the univariate hypotheses, after adjusting for Type I

error. (Confidence intervals also can be reported in Step 1.) In Step 2, as is the case for Robinson and Levin's (1997) model, if the *p*-value is less than the adjusted alpha level (i.e., statistical significance), then the appropriate effect-size index (i.e., practical significance) is reported, alongside their confidence intervals (i.e., theoretical significance). The Binomial Test of Result Direction begins the third step in the process. Specifically, if the Binomial test yields a *p*-value greater than .05, then no effect-size index would be reported, and the researcher would conclude that there was no statistically significant trend; on the other hand, if the Binomial test yields a *p*-value less than .05 (i.e., statistically significance), then the analyst would report and interpret the proportion-observed-above chance index. (Confidence intervals, again, could supplement the Binomial Test of Result Direction and any associated effect size.)

Insert Figure 1 about here

As can be seen, the Three-Step Model for testing multiple univariate null hypotheses helps to reduce the probability of Type A error (see Table 1) relative to the Robinson and Levin's (1997) Two-Step Model. A desirable feature of the Binomial Test of Result Direction is the fact that this represents a non-parametric test of significance, and thus has less stringent assumptions than is the case for parametric null hypothesis tests of significance. An additional appeal of the Binomial test of significance is that it provides exact *p*-values.

It should be noted that the Binomial Test of Result Direction is only appropriate when five or more univariate null hypotheses are of interest. This is because, whereas the probability of observing 5 successes in 5 trials (i.e., hypothesis test) under the null hypothesis of equal proportions is .031, the probability of observing 4 successes in 4 trials is .062. Table 2 presents the number of successes needed to obtain a *p*-value less than .05 (i.e., critical value) using the Binomial Test of Result Direction. In this table, the number of hypotheses tested (Range = 5-25) is presented in the first column, whereas the critical value is specified in the second column. The final column represents the probability of obtaining a number of successes as extreme or more extreme than the critical value. Thus, for example, if 8 univariate null hypothesis significance tests were conducted, then either 7 or 8 findings in the same direction would yield a statistically significant trend because the probability of obtaining 7 or more successes in 8 trials is .035.

Insert Table 2 about here

Heuristic Example

Recently, Elbedour, Onwuegbuzie, and Stills (2001) were interested in determining whether it makes a difference to children from the Bedouin-Arab community in Israel if children live in monogamous or polygamous families with respect to a myriad of school performance and psychological factors. These researchers administered a Hebrew version of the Teacher Report Form (TRF) of the Child Behavior

Checklist (CBCL; Achenbach, 1991). This Hebrew version was developed by Auerbach, Goldstein, and Elbedour (2000), using the translation-back translation method. The CBCL is the most commonly used measure of social competence and behavioral problems. The CBC contains 8 subscales (e.g., anxiety, delinquency, aggression) that can be collapsed into an internalizing or externalizing scale. Participants, who consisted of 153 children from one-wife families and 102 children from two-wife families, also were administered 10 measures of academic achievement and 11 measures of teacher ratings.

When Elbedour et al. compared (in a univariate manner) the two sets of children with respect to the 10 achievement measures, after applying the Bonferroni adjustment, they found that all of the *p*-values exceeded the Bonferroni-adjusted alpha value of .006. Thus, if the investigator had used the two-step method, they would have concluded that there was no difference between the two groups with respect to school performance, even though 8 of the 10 p-values were less than .05. Not being able to reject any of the 10 null hypotheses prevented them from reporting effect sizes under the Two-Step rule. Yet, what was extremely compelling about their achievement data was the fact that the direction of the difference was the same for all 10 measures. Specifically, for all 10 measures, the one-wife children outperformed the two-wife children. Thus, they applied the third step of the Three-Step model proposed above, and conducted a Binomial Test for Result Direction, which yielded a *p*-value of ..001. Thus, although the researchers were not able to state that one-wife children obtained statistically significantly higher scores on any of the achievement variables, they were able to conclude the following:

> In the present study, the direction of the difference was the same for all 10 achievement measures. Specifically, for all 10 indicators, one-wife children outperformed the two-wife children. This result is extremely compelling because under the null hypothesis that the one- and two-wife children were equivalent across all levels of academic performance, the probability of all 10 outcomes being in favor of one-wife children was .001. Thus, the Binomial Test of Result Direction revealed that the consistency with which children from one-wife families outperformed those from two-wife families was statistically significant ($p < .001$). The magnitude pertaining to this trend, as measured by the difference between the observed proportion and the proportion expected by chance, was 1.00 - 0.50 = 0.50 (Onwuegbuzie, 2001). This magnitude can be expressed as a proportion-observed-above-chance effect size measure of 100(1.00 - .05)/0.50 = 100%, which is the maximum value for this Binomial effect size index. (p. 25)

Summary and Conclusions

Much debate still prevails regarding the use of statistical significance tests in empirical studies. Whereas some researchers believe that statistical significance testing should be banned completely (e.g., Bakan, 1966; Cahan, 2000; Carver, 1978, 1993; Cohen, 1994, 1997; Guttman, 1985; Loftus, 1996; Meehl, 1967, 1978; Nix & Barnette, 1998; Rozeboom, 1960; Schmidt, 1992; 1996; Schmidt & Hunter, 1997), others contend that statistical significance tests should play a role in statistical analyses (Abelson,

1997a, 1997b; Frick, 1996; Levin, 1993, 1998; McLean & Ernest, 1998; Mulaik, Raju, & Harshman, 1997; Onwuegbuzie & Daniel, in press-a, in press-b). The former group tend to argue that effect sizes should replace *p*-values completely. However, in the current paper, 10 flaws associated with effect-size reporting were described in detail, thereby refuting this contention.

Researchers who believe that statistical significance tests should remain tend to to advocate that *p*-values and effect sizes complement one another. However, among this group, debate exists as to how these indices should be combined, ranging from those who maintain that effect sizes always should be reported (e.g., Thompson, 1996), to those who believe that effect sizes only should be reported if the observed finding is statistically significant (e.g., Robinson & Levin, 1997). The concepts of Type A error (i.e., over-interpreting large effect sizes in the presence of statistically non-significant findings) and Type B error (over-interpreting a small effect size in the presence of a statistically significant finding) were introduced to refute the use of *p*-values and effect sizes in isolation. Rather, Robinson and Levin's (1997) Two-Step procedure was advanced, whereby researchers first objectively evaluate the probability of an observed effect (i.e., statistical significance) and, if and only if it can be concluded that the underlying finding is too improbable to be due to chance, then its magnitude or effect size (i.e., practical significance) is assessed. It was contended that this technique is extremely useful when conducting null hypothesis significance tests that have sufficient statistical power.

Unfortunately, evidence was provided that the Two-Step method was inadequate at best, and misleading at worst, when multiple tests of univariate hypotheses are of interest due, to the reduction in statistical power stemming from adjustments for Type I error that are needed. In such cases, a Three-Step Method was proposed to replace the Two-Step model. This three-step method involves a Binomial Test of Result Direction at its third stage. A heuristic example was provided to illustrate the utility of this approach.

The Binomial Test of Direction proposed in this paper has other uses. For example, it can be used to analyze correlation matrices (e.g., as part of a multi-trait multimethod analysis) to determine how many relationships among instrument subscales are in a statistically significantly consistent direction (i.e., all or most in either the positive or negative direction). Also, it could be used in discriminant analysis, canonical analysis, and factor analysis to determine either whether the loadings are in a statistically significantly consistent direction, or to test whether the proportion of loadings that are practically significant (e.g., greater than .3 or .5) are above what could be expected by chance. Further, this test could be used in structural equation modeling to determine whether the proportion of significant paths are above what could be expected by chance.

Finally, the binomial test perhaps could be used in meta-analyses to determine whether the number of effects across studies that are statistically significant is above what could be expected by chance or whether the number of effect sizes above an *a priori* effect size across the selected studies represents chance. As such, it is hoped

that the Binomial Test of Result Direction in particular and the Three-Step Model in general that have been proposed here are subjected to scientific scrutiny, but one that is as balanced as it is rigorous.

References

Abelson, R. P. (1997a). The surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*, 12-15.

Abelson, R. P. (1997b). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117-141). Mahwah, NJ: Erlbaum.

Achenbach, T. M. (1991). *Manual for the Teacher's Report Form and 1991 Profile*. Burlington: University of Vermont, Department of Psychiatry.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (1999). *Null hypothesis testing in ecological studies. Problems, prevalence, and alternative*. Manuscript submitted for publication.

Auerbach, J. G., Goldstein, E., & Elbedour, S. (2000). Behavior problems in Bedouin elementary schoolchildren. *Transcultural Psychiatry, 37*, 229-241.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423-437.

Barnette, J. J., & McLean, J. E. (1999, November). *Empirically based criteria for determining meaningful effect size*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Point Clear, AL.

Barnette, J. J., & McLean, J. E. (2000, April). *Use of significance test as protection against spuriously high standardized effect sizes: Introduction of the protected effect size*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-536.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37*, 325-335.

Boring, E. G. (1919). Mathematical vs. scientific importance. *Psychological Bulletin, 16*, 335-338.

Cahan S. (2000). Statistical significance is not a "Kosher Certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results. *Educational Researcher, 29*(1), 31-34.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D.T., & Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61, 287-292*.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A

review. *Journal of Abnormal Psychology, 65*, 145-153.
Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: McGraw-Hill.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: John Wiley.
Cohen, J. (1994). The earth is round (r < .05). *American Psychologist, 49*, 997-1003.
Cohen, J. (1997). The earth is round (r < .05). In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 21-35). Mahwah, NJ: Erlbaum.
Cortina, J. M., & Dunlap, W. P. (1997). Logic and purpose of significance testing. *Psychological Methods, 2*, 161-172.
Cruise, R. J., & Wilkins, E. M. (1980). *STARS: Statistical Anxiety Rating Scale.* Unpublished manuscript, Andrews University, Berrien Springs, MI.
Daniel, L.G., & Onwuegbuzie, A.J. (2000, November). *Toward an extended typology of research errors.* Paper presented at the annual conference of the Mid-South Educational Research Association, Bowling Green, KY.
Darr, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. *American Psychologist, 42*, 145-151.
Elbedour, S., Onwuegbuzie, A. J., & Stills, A. (2001). *Behavioral problems and scholastic adjustment among Bedouin-Arab children from polygamous and monogamous marital family structures.* Manuscript submitted for publication.
Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*, 75-98.
Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94*, 275-282.
Fisher, R.A. (1925/1941). *Statistical methods for research workers* (84th ed.) Edinburgh, Scotland: Oliver & Boyd. (Original work published in 1925).
Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1*, 379-390.
Guttman, L. B. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis, 1*, 3-10.
Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997, Eds.). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.
Hogarty, K. Y., & Kromrey, J. D. (2001, April). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the annua meeting of the American Educational Research Association, Seattle, WA.
Huck, S.W. (2000). *Reading statistics and research* (3rd ed.). New York: Addison Wesley Longman.
Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*, 3-7.
Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B.,

Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Kirk, R. E. (1996). Practical significance. A concept whose time as come. *Education and Psychological Measurement, 56*, 746-759.

Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta analysis effect size calculation. *Psychological Bulletin, 91*, 404-412.

Krantz, D. H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association, 94*(448), 1372-1382.

Levin, J. R. (1993). Statistical significance testing from three perspectives. *Journal of Experimental Education, 61*, 378-381.

Levin, J. R. (1998). What if there were no more bickering about statistical significance tests? *Research in the Schools, 5*, 43-53.

Levin, J. R., & Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychological Review, 11*, 143-155.

Levin, J. R., & Robinson, D. H. (2000). Statistical hypothesis testing, effect-size estimation, and the conclusion coherence of primary research studies. *Educational Researcher, 29*(1), 34-36.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychology*, 5, 161-171.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361-365.

McLean, J. E. (1995). *Improving education through action research: A guide for administrators and teachers*. Thousand Oaks, CA: Corwin Press.

McLean, J. E., O'Neal, M. R., & Barnette, J. J. (2000). *Are all effect sizes created equal?* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

McLean, J. E., & Ernest, J. M. (1998). The role of statistical significance testing in educational research. *Research in the Schools, 5*, 15-22.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and a place for significance testing. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65-115). Mahwah, NJ: Erlbaum.

Neyman, J., & Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 29A*, Part I: 175-240; part II 263-294.

Nix, T. W., & Barnette, J. J. (1998). The data analysis dilemma: Ban or abandon. A

review of null hypothesis significance testing. *Research in the schools, 5*, 3-14.

Onwuegbuzie, A.J. (2001a, April). *Effect sizes in qualitative research: A prolegomenon.* Paper presented at the annual conference of the American Educational Research Association (AERA), Seattle, WA.

Onwuegbuzie, A.J. (2001b). *Towards a framework for comprehensive reporting of empirical findings: The role of statistical significance, theoretical significance, practical significance, and clinical significance.* Unpublished manuscript, Howard University, Washington, DC.

Onwuegbuzie, A.J. (in press-a). Common analytical and interpretational errors in educational research: an analysis of the 1998 volume of the British Journal of Educational Psychology. *Research for Educational Reform.*

Onwuegbuzie, A.J. (in press-b). Expanding the framework of internal and external validity in quantitative research. *Research in the Schools.*

Onwuegbuzie, A.J. (in press-c). Positivists, post-positivists, post-structuralists, and post-modernists: Why can't we all get along? Towards a framework for unifying research paradigms. *Education.*

Onwuegbuzie, A.J., & Daniel, L.G. (2000, November). *Reliability generalization: The importance of considering sample specificity, confidence intervals, and subgroup differences.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Onwuegbuzie, A.J., & Daniel, L.G. (2001, April). *Indices of score reliability and their applications.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Onwuegbuzie, A.J., & Daniel, L.G. (in press-a). Uses and misuses of the correlation coefficient. *Research in the Schools, 9*(1).

Onwuegbuzie, A.J., & Daniel, L.G. (in press-b). Typology of analytical and interpretational errors in quantitative and qualitative educational research. *Current Issues in Education.*

Onwuegbuzie, A.J., Daniel, L.G., & Roberts, J.K. (2001, November). *A proposed new "what if" reliability analysis for assessing the statistical significance of bivariate relationships.* Paper to be presented at the annual conference of the Mid-South Educational Research Association, Little Rock, AR.

Roberts, J.K., & Onwuegbuzie, A.J. (2000, November). *Alternative approaches for interpreting alpha with homogeneous subsamples.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Bowling Green, KY.

Roberts, J.K., Onwuegbuzie, A.J., & Eby, R. (2001, April). *Alternative approaches for interpreting alpha with homogeneous subsamples: The introduction of a new measure of homogeneous alpha.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Robinson, D. H., & Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26*(5), 21-26.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test.

Psychological Bulletin, 57, 416-428.

Rozeboom, W. W. (1997). Good science is abductive, not hypothetic-deductive. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 335-392). Mahwah, NJ: Erlbaum.

Schmidt, F. L. (1992). What do data really mean? *American Psychologist, 47*, 1173-1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L.L. Harlow, S.A. Mulaik, & J.H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Erlbaum.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316.

Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education, 61*(4), 350-360.

Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education, 61, 293-316.*

Tatsuoka, M.M. (1973). Multivariate analysis in educational research. In F.N. Kerlinger (Ed.), *Review of Research in Education* (pp. 273-319). Itasca, IL: Peacock.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70*, 434-438.

Thompson, B. (1993). The use of statistical significance research: Bootstrap and other alternatives. *The Journal of Experimental Education, 61*, 361-377.

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30.

Thompson, B. (1999, April). Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap. Invited address presented at the annual meeting of the American Educational Research Association, Montreal [On-line]. Available: *http://acs.tamu.edu/~bbt6147/aeraad99.htm*

Tryon, W. W. (1998). The inscrutable null hypothesis. *American Psychologist, 53*, 796.

Weems, G.H., & Onwuegbuzie, A.J. (in press). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development.*

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika, 61*, 165-170.

Table 1

Possible Outcomes and Conclusion Validity When Null Hypothesis Significance Testing and Effect Sizes are Combined

		Effect Size	
		Large	Small
p-value	Not statistically significant	Type A error	Conclusion Validity
	Statistically significant	Conclusion Validity	Type B error

Table 2

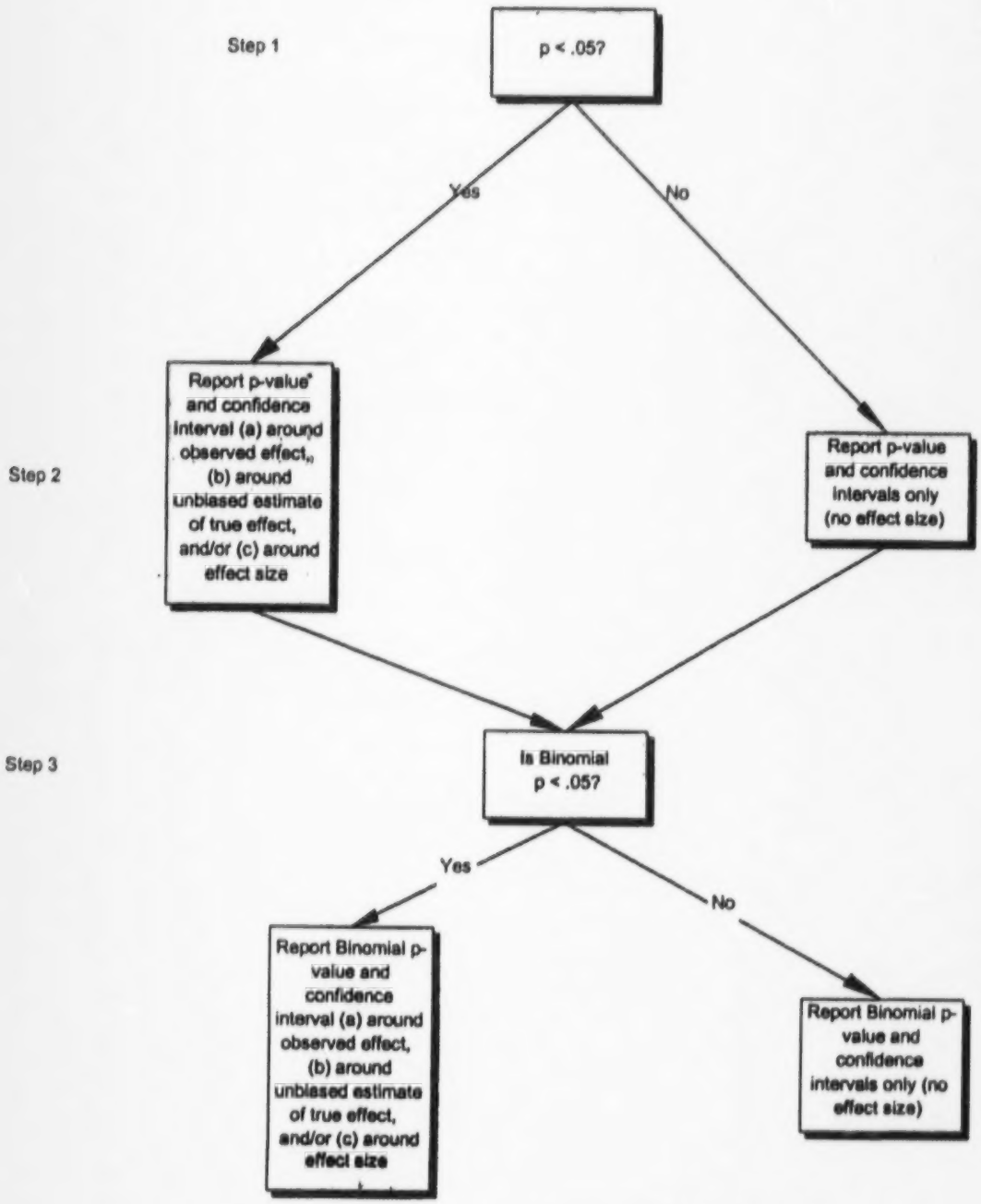
Number of Successes Needed to Obtain a Statistically Significant Trend (i.e., Critical Value) Using the Binomial Test of Result Direction

Number of Null Hypotheses Tested	Critical Value[1]	*p*-value[2]
5	5	.0310
6	6	.0160
7	7	.0080
8	7	.0350
9	8	.0200
10	9	.0110
11	9	.0327
12	10	.0193
13	10	.0461
14	11	.0287
15	12	.0180
16	12	.0384
17	13	.0245
18	13	.0481
19	14	.0318
20	15	.0210
21	15	.0392
22	16	.0262
23	16	.0466
24	17	.0320
25	18	.0216

[1] The critical value represents the number of successes (i.e., observed findings in the same direction) needed to obtain a *p*-value less than .05 using the Binomial Test of Result Direction.

[2] The *p*-value represents the probability of obtaining a number of successes as extreme or more extreme than the critical value

Figure 1. *Three-step method for testing multiple univariate null hypotheses.*


Step 1
p < .05?
Yes
No
Step 2
Report p-value and confidence interval (a) around observed effect, (b) around unbiased estimate of true effect, and/or (c) around effect size
Report p-value and confidence intervals only (no effect size)
Step 3
Is Binomial p < .05?
Yes
No
Report Binomial p-value and confidence interval (a) around observed effect, (b) around unbiased estimate of true effect, and/or (c) around effect size
Report Binomial p-value and confidence intervals only (no effect size)

BEST COPY AVAILABLE